

A satellite image of Earth's oceans and clouds, showing swirling patterns of white clouds over dark blue and brownish ocean and land masses. The image is used as a background for the title text.

Facilitating New Opportunities for Data Users via NOAA's Big Data Project

Dr. Edward J. Kearns
Chief Data Officer

National Oceanic and Atmospheric Administration

NOAA Satellite Conference Big Data Panel 17 July 2017

Acknowledgements

Many thanks to:

- BDP Core Team: Andy Bailey, Shane Glass, Jeff de la Beaujardiere, Tony LaVoi, Jay Morris, Derek Parks
- NOAA: Brian Eiler*, Zach Goldstein, Dave Michaud, Glenn Tallia, Derek Hanson, Kate Abbott, Amy Gaskins*, Alan Steremberg*, Maia Hansen*, Steve Ansari, Steve Del Greco*, Brian Nelson, Carlos Rivero*, Ken Casey, Rich Baldwin, Ed Clark, Brian Cosgrove, Steve Volz, Mark Paese, Donna McNamara, Chris Sisko, Nathan Wilson, Mark Brady*, Renata Lana
- NC State University / CICS-NC: Otis Brown, Scott Wilkins, Jon Brannock, Lou Vazquez, Scott Stevens, Paula Hennon*, Andrew Buddenberg, Angel Li

NOAA's Big Data Collaborators and their partners (not an all inclusive list)

- Amazon: Jed Sundwall, Arial Gold*, Jeff Layton, Joe Flasher
- Microsoft: Sam Khoury, Sid Krishna, Shannon Murphy
- Google: Will Curran, Matt Hancher, Eli Bixby, Tino Tereshko, Amy Unruh, Tanya Shastri, Ossama Alami, Valliappa "Lak" Lakshmanan^, Mike Hamberg
- Open Commons Consortium: Walt Wells, Maria Patterson, Zac Flamig
- Unidata: Mohan Ramamurthy, Jeff Weber
- IBM: James Stevenson, Stefani Jones, Mary Glackin, Peter Neilley, John Aviles
- The Climate Corporation: Adam Pasch

Why is NOAA so interested in Partnerships for Open Data?

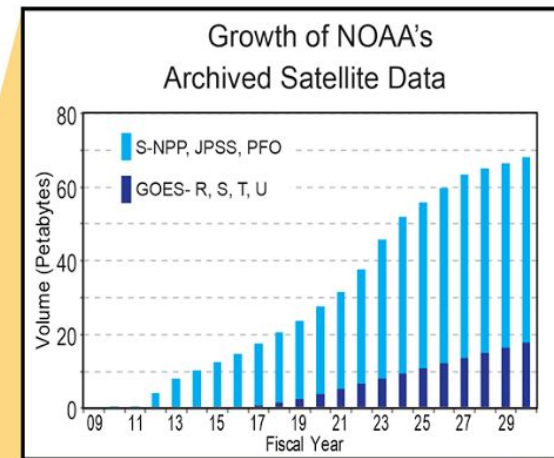
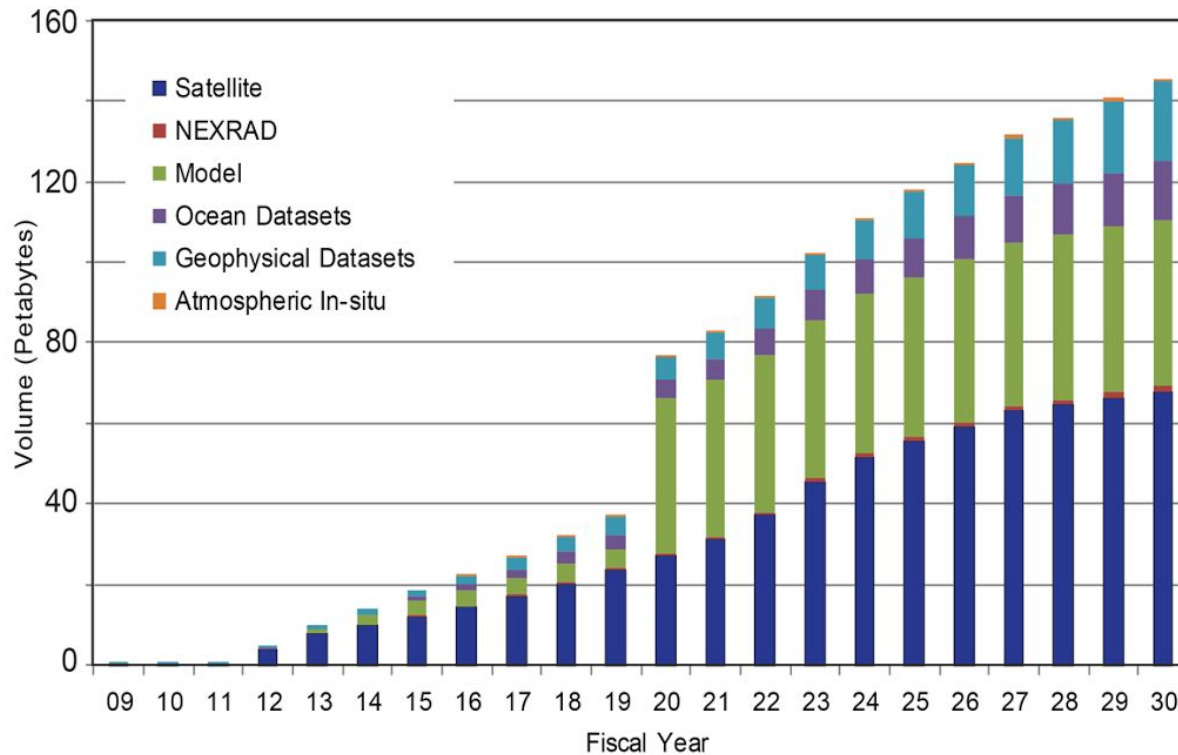
- NOAA's **full and open data** are **increasingly popular and valuable**.
- NOAA struggles to keep up with **increasing public demand**
 - Budgets for additional data access capacity and capabilities: Flat
 - NOAA Costs for data access: Rapidly increasing
- NOAA wants to learn about collaborative solutions
 - Promote use, democratize data access
 - Utilize new technologies
 - Enable new economic opportunities for partners.

Improve Accessibility to NOAA's Open Data

Why is NOAA interested in this?

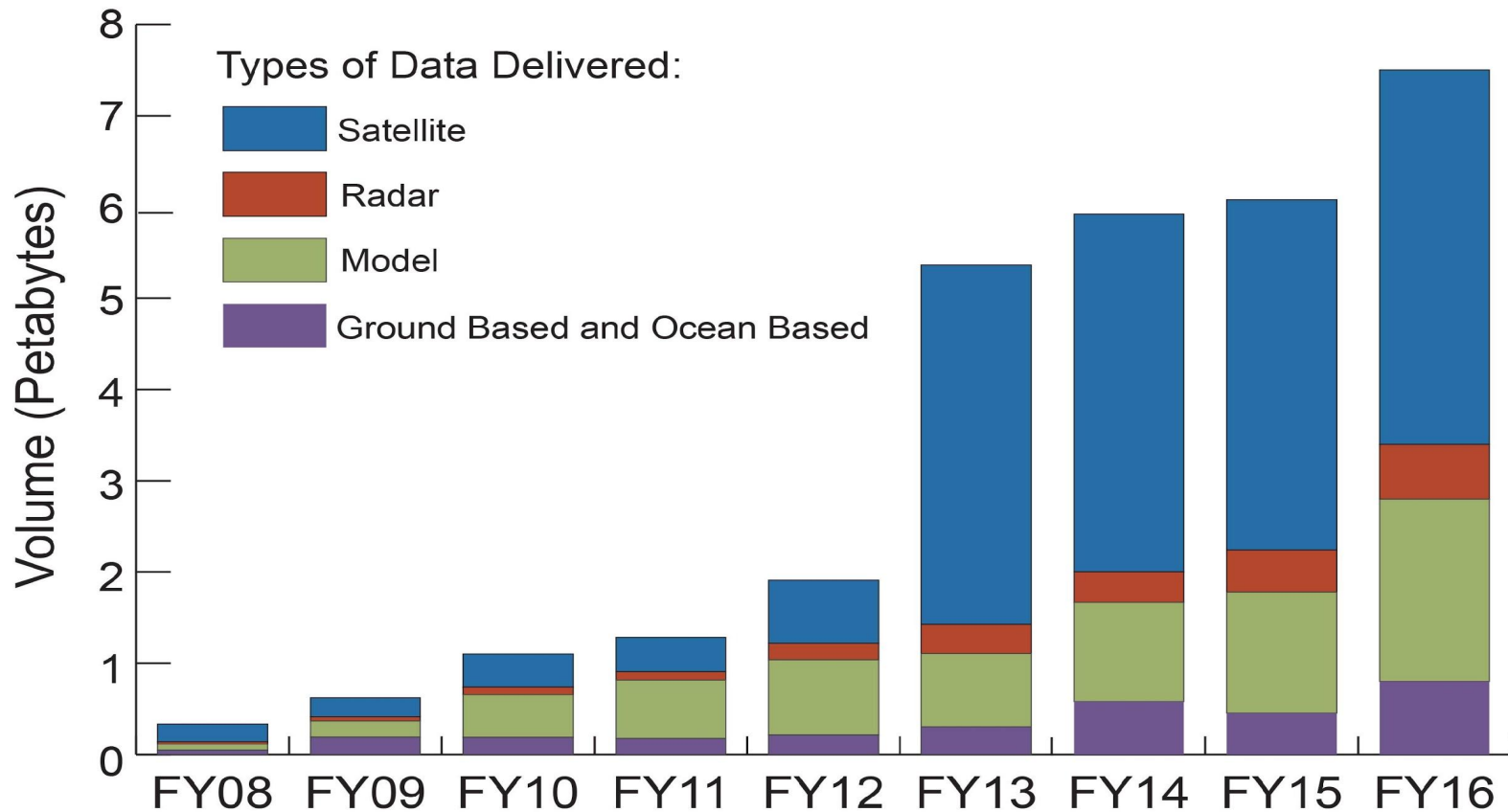
Projections for NOAA Archived Data

Growth of NOAA's Archive



Why is NOAA interested in this?

NOAA Archived Data Access by Volume

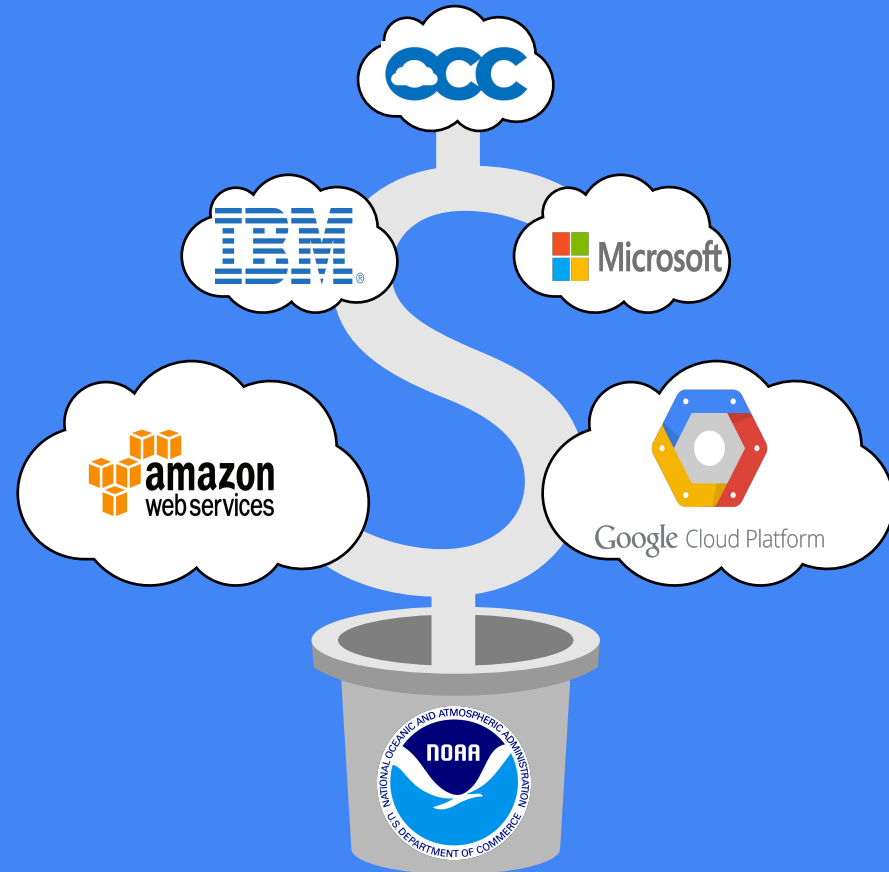


The Big Data Project

A Business Experiment

Keys

- Bring users to the data
 - Not “just” about access
- CRADAs - research activity (2015)
- NOAA’s open data
- NOAA’s subject matter expertise
- Industry’s infrastructure expertise
- Level playing field
 - No privileged access
- Democratization of NOAA data
 - New opportunities for business



Leverage the value of NOAA's data to increase their utilization

Big Data Project Methodology

01

Business Discovery

CRADA Collaborators & any Third-Party Partners work together to identify datasets of interest & develop business cases

02

Initial Technical Discussion

Develop a strategy for data delivery from NOAA to BDP Collaborators

03

In-Depth Data Discussions

Engage NOAA SMEs, BDP Collaborators for technical interchanges

04

Product Development

Collaborators and their Partners create services

- ◆ Develop markets & financial opportunities based on NOAA data
- ◆ Generate revenue and profits

05

Augmented NOAA Services

NOAA continues all of it's existing data services

- No interruption of existing services to customers, but new options
- BDP activities are an augmentation of



NOAA Big Data Project Data Access Strategy

Collaborate with Industrial Partners to Learn

Augment



Add
Capabilities

Amplify

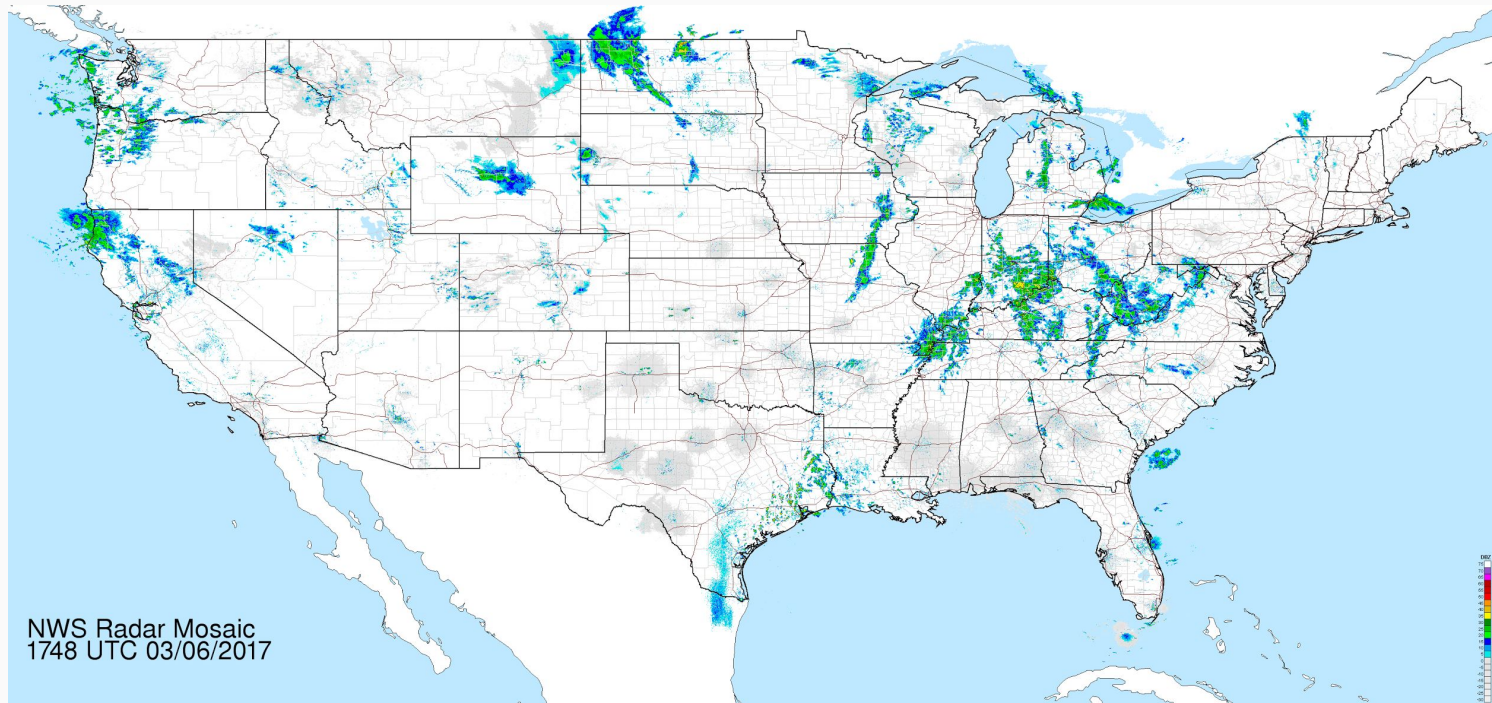


Add
Capacity

Example BDP Success Story

NEXRAD Radar Data : 1991- Present

- Entire NWS NEXRAD Level 2 Archive (300 TB) was transferred from NCEI to AWS, OCC (2015-17), Microsoft, and Google



Example BDP Success Story

NEXRAD Level 2 Radar Data on AWS

Data Usage

Increased 2.3X



Decreased 50%



Archive Server Load

Ansari et al., 2017. Unlocking the potential of NEXRAD data through NOAA's Big Data Partnership
<http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-16-0021.1>

Example BDP Success Story

NEXRAD Level 2 Radar Data on AWS

NOAA Wins



■ AWS

■ NCEI

AWS?



End User Wins



■ AWS Job Time ~days
■ Through NCEI ~Years

OCC NEXRAD Access

<http://edc.occ-data.org/nexrad/>

NOAA NEXRAD WSR-88D

Get Data

How to get NEXRAD data from the
OCC Environmental Data Commons

Python | Jupyter

Work with NEXRAD data using
Python and Jupyter

Signpost ID Service

How to use the Environmental Data
Commons ID Service



- Contact: info@occ-data.org || © 2017 O
- Powered by [Hugo](#) and the [Kube theme](#)

Google NEXRAD Access

<https://cloud.google.com/blog/big-data/2017/06/visualization-and-large-scale-processing-of-historical-weather-radar-nexrad-level-ii-data>



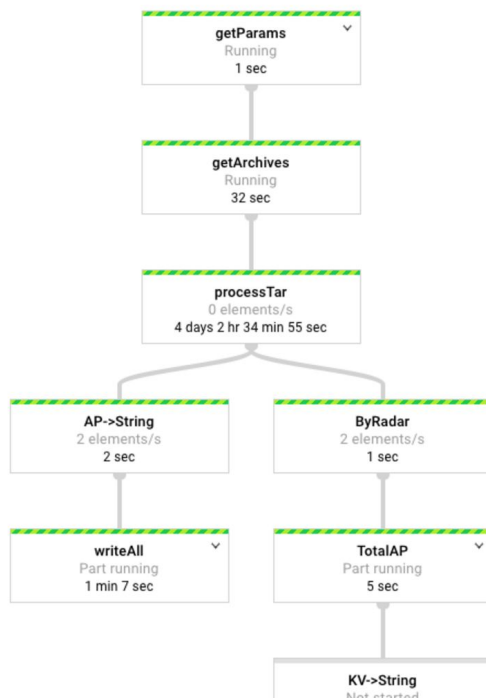
Why Google Products Solutions Launcher Pricing Customers Documentation Support Partners

Sample program to do large-scale analysis

While you can work with individual volume scans as shown above, one key benefit of having all the NEXRAD data immediately available on a public cloud is the ability to analyze long time periods of data at scale. Thanks to GCP's "serverless" approach to infrastructure, it's possible to do data processing, data analysis and machine learning without having to manage low-level resources.

[Cloud Dataflow](#), GCP's fully-managed service for stream and batch processing, allows you to write a data processing and analysis pipeline that will be executed in a distributed manner. The pipeline will autoscale different steps to run on multiple machines in a fault-tolerant way.

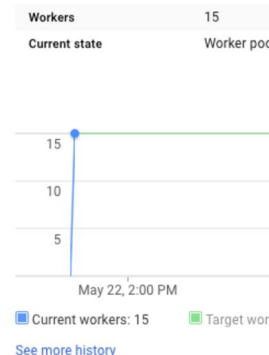
As of June 15, 2017



Job summary

Job name	appipeline-vlakshmanan-0522205052-597f064b
Job ID	2017-05-22_13_50_54-9553301786727587053
Job status	Running
SDK version	Google Clo Java 2.0.0-
Job type	Batch
Start time	May 22, 20
Elapsed time	53 min 26 s

Autoscaling



Google Cloud Platform

Select a project

Storage

Browser

UPLOAD FILES

UPLOAD FOLDER

CREATE FOL

Filter by prefix...

Buckets / gcp-public-data-nexrad-12 / 2015 / 04 / 01 / KABR

<div><div></div></div> <div>Name</div>	Size
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401000000_20150401005959.tar</div> <div>17.6 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401010000_20150401015959.tar</div> <div>25.45 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401020000_20150401025959.tar</div> <div>26.66 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401030000_20150401035959.tar</div> <div>28.64 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401040000_20150401045959.tar</div> <div>29.91 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401050000_20150401055959.tar</div> <div>30.26 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401060000_20150401065959.tar</div> <div>29.38 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401070000_20150401075959.tar</div> <div>31.68 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401080000_20150401085959.tar</div> <div>23.21 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401090000_20150401095959.tar</div> <div>21.12 MB</div>	
<div><div></div></div> <div>NWS_NEXRAD_NXL2DP_KABR_20150401100000_20150401105959.tar</div> <div>20.54 MB</div>	

Google Cloud Platform Example

The screenshot shows the Google Cloud Platform documentation page for the NOAA Global Historical Climatology Network (GHCN) Weather Data. The page is titled "NOAA Global Historical Climatology Network Weather Data" and is part of the BigQuery documentation. It includes a left sidebar with a "Resources" section listing various datasets like "NOAA GHCN Weather", "NOAA GSOD Weather", and "NYC 311 Service Requests". The main content area describes the GHCN dataset, which is an integrated database of climate summaries from land surface stations across the globe. It mentions that two GHCN datasets are available in BigQuery: the GHCN-D (daily) and the GHCN-M (monthly). The data includes information from more than 20 sources, including some data from every year since 1763. There are links to "GO TO NOAA GHCN DATASET (DAILY)" and "GO TO NOAA GHCN DATASET (MONTHLY)". A "Sample queries" section provides an example of an SQL query to find weather stations close to a specific location, such as Chicago. The query is:

```
SELECT id,
```

- **1.2 PBs of climate and weather data accessed through Google BigQuery, from Jan-Apr 2017**

- Without “trying” - not advertised yet
- Joins, joins, joins
- 30-100x of NOAA deliveries in that time

- Images in Google Earth Engine

- GOES-16 (June 2017)
- National Water Model data
- Weather and Climate model output
- Climate data records

Big Data Project Collaborators' Data Offerings

- **Amazon Web Services (AWS)**
 - <https://aws.amazon.com/noaa-big-data/>
- **Google Cloud Platform**
 - <https://cloud.google.com/bigquery/public-data/>
- **IBM**
 - <https://noaa-crada.mybluemix.net/node/32>
- **Microsoft Azure**
 - Public Services TBD
- **Open Commons Consortium (OCC)**
 - <http://edc.occ-data.org/>

Big Data Project and Open Data Challenges

- How well do we understand the Big Data market?
 - Importance of 3rd parties in understanding the market values
 - Will the market create and shape the services it needs?
- Efficiencies of Use and the Marginal Cost of Distribution
 - Cloud Computing Platform versus a Distribution Network
- How to best transfer and steward many large, complex datasets?
 - How to ensure data integrity and authenticity?
 - Real-time, e.g. satellites, weather observations, coastal data
 - Retrospective, e.g. climate models and observations, fisheries
- Next Data Sets to bring into this demonstration project
 - **GOES-16**, National Water Model, CFS/NMME, GFS/HRRR, others...

Big Data Project Opportunities

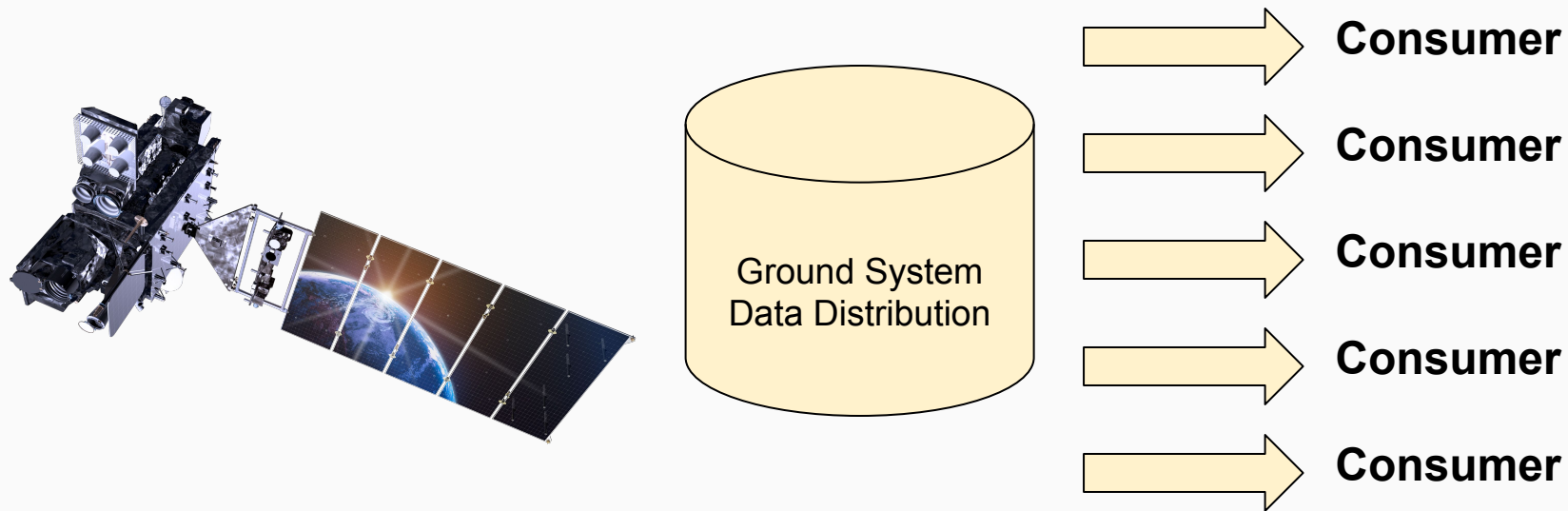
- Enhanced distribution of NOAA's open data
- Reduced level of effort for public data access
 - Don't have to move the data to use them
 - Use this experience to inform future dissemination strategies
- High Level of Service to customers
 - Is there value in higher levels of service?
- This is not *just* about open data access
 - Can accelerate data utilization...
 - ...and thus societal impacts and business opportunities

GOES-16 Satellite Products and Services

- Please see our NESDIS leadership and their staffs for specific information on GOES-16 products and services
 - Steve Volz
 - Mark Paese
 - Karen St. Germain
 - Vanessa Griffin
- The Big Data Project (BDP) is a demonstration effort and business experiment and is not an operational function.
 - We wish to learn from the BDP experiment to help inform future NOAA and NESDIS decisions on open data distribution to our many users.

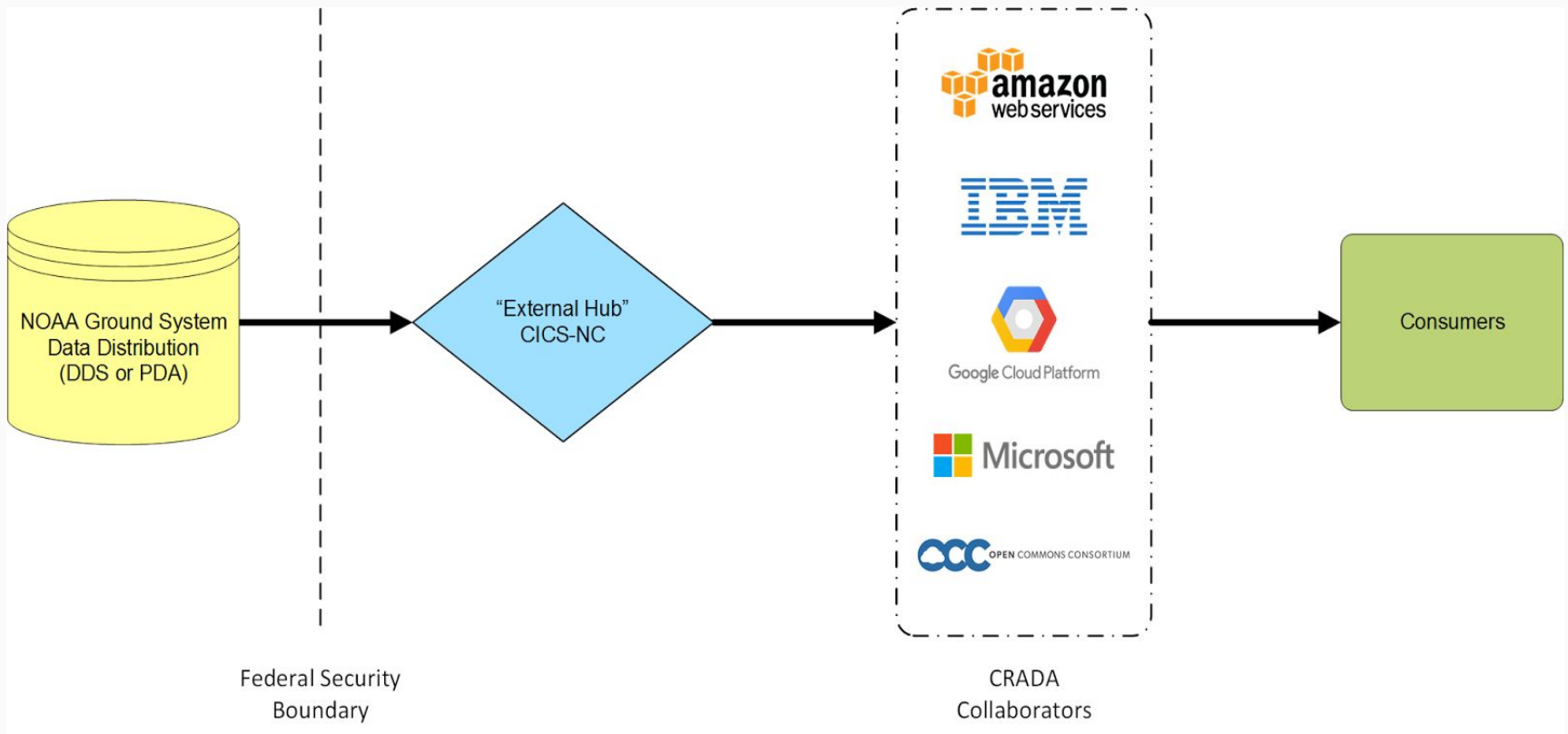
Traditional Satellite Data Internet Access Strategy

One-to-One Model



Big Data Project Satellite Data Access Demo Activity

One-to-Many Model



GOES-16 BDP Demo Live as of July 12, 2017: Initial Distribution Statistics

- The BDP is partnering with the Cooperative Institute for Climate and Satellites - North Carolina (CICS-NC) to provide feeds of the GOES-16 data from the NOAA Ground System (as an authorized user) to the BDP CRADA Collaborators.
- CICS-NC is offering 5 validated feeds to the BDP Collaborators
 - timing - as fast as they appear at NOAA distribution point
 - single bounce of data through CICS-NC systems, w/checksums
 - minimizes load on NOAA's operational systems and networks
- **Observed additional latencies from CICS-NC transfer mechanism**
 - **From NOAA Ground System to BDP Collaborator platforms**
 - **Maximum additional latency: 2 to 3 min (full disk ABI, Band 2)**
 - **Typical Range of additional latency: 30 sec - 3 min**

BDP Collaborators' GOES-16 Data Platforms

- **AWS**
 - <https://aws.amazon.com/public-datasets/goes/>
- **Google Cloud Platform**
 - Public Services TBD
- **IBM**
 - Public Services TBD
- **Microsoft Azure**
 - Public Services TBD
- **Open Commons Consortium (OCC)**
 - <http://edc.occ-data.org/goes16/>

AWS GOES-16

<https://aws.amazon.com/public-datasets/goes/>



Products ▾

Solutions

Pricing

Software

Support

Customers

Partners

Enterprises

More ▾

English ▾

My Account ▾

RELATED LINKS

[Big Data on AWS](#)

[Open Data on AWS](#)

[AWS Programs for Research and Education](#)

GOES-R Series on AWS

Data from NOAA's GOES-R series satellite is available on Amazon S3. The National Oceanic and Atmospheric Administration (NOAA) operates a constellation of Geostationary Operational Environmental Satellites (GOES) to provide continuous weather imagery and monitoring of meteorological and space environment data for the protection of life and property across the United States. GOES satellites provide critical atmospheric, oceanic, climatic and space weather products supporting weather forecasting and warnings, climatologic analysis and prediction, ecosystems management, safe and efficient public and private transportation, and other national priorities.

The satellites provide advanced imaging with increased spatial resolution, 16 spectral channels, and up to 1 minute scan frequency for more accurate forecasts and timely warnings.

The real-time feed and full historical archive of original resolution Advanced Baseline Imager (ABI) radiance data (Level 1b) and full resolution Cloud and Moisture Imager (CMI) products (Level 2) are freely available on Amazon S3 for anyone to use. Currently, GOES-16 data is at provisional status. Please see details of the data maturity [here](#).

Accessing GOES Data on AWS


While the GOES-16 ABI L1b and CMI data have reached provisional validation, please keep in mind that since GOES-16 satellite has not been declared operational, its data are still considered preliminary and undergoing testing.

The availability of GOES-R Series on AWS data is the result of the NOAA Big Data Project (BDP) to explore the potential benefits of storing copies of key observations and model outputs in the Cloud to allow computing directly on the data without requiring further distribution. Such an approach could help form new lines of business and economic growth while making NOAA's data more easily accessible to the American public.

This page includes information on data structure; you can find much more detailed information about GOES-R Series data from NOAA [here](#).

AWS GOES-16

<https://aws.amazon.com/public-datasets/goes/>

Products ▾ Solutions Pricing Software Support Customers Partners Enterprises More ▾English ▾ My Account ▾

```
aws s3 ls noaa-goes16

aws s3 cp s3://noaa-goes16/<Product>/<Year>/<Day of Year>/<Hour>/<Filename>
```

Subscribing to GOES Data Notifications

We have set up public [Amazon SNS](#) topics that create a notification for every new object added to the Amazon S3 buckets for GOES on AWS. To start, you can subscribe to these notifications using [Amazon SQS](#) and [AWS Lambda](#). This means you can automatically add new real-time and near-real-time GOES data into a queue or trigger event-based processing if the data meets certain criteria such as geographic location.

The ARN for the PDA feed is **arn:aws:sns:us-east-1:123901341784:NewGOES16Object**.

About the Data

Source	National Oceanic and Atmospheric Administration
Category	Earth Science, Sensor Data, Natural Resource, Meteorological
Format	netCDF v4
License	There are no restrictions on the use of this data.
Storage Service	Amazon S3
Location	s3://noaa-goes16 in us-east-1 region
Update Frequency	New data is added as soon as it's available

Earth on AWS Cloud Credits for Research

Educators, researchers and students can apply for free promotional credits to take advantage of Public Datasets on AWS. If you have a research project that could take advantage of GOES data on AWS, you can apply for [Earth on AWS Cloud Credits for Research](#).

Google GOES-16

No URL provided yet.

OCC's Environmental Data Commons

<http://edc.occ-data.org/>

The OCC Environmental Data Commons

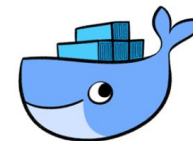
Repository for environmental public data sets of scientific interest, hosted as
part of the Open Science Data Cloud Ecosystem



[GOES 16](#)



[NEXRAD](#)



[Tools | Notebooks](#)

OCC GOES-16 Resources

<http://edc.occ-data.org/goes16/>

GOES-16 / GOES-R

Get Data

How to get GOES-16 data from the
OCC Environmental Data Commons

Using Python to Explore GOES-16 Data

Working with GOES-16 data using
Python and Jupyter

Manipulating GOES-16 Data with GDAL

Using GDAL to Work with GOES-16
NetCDF Data



- Contact: info@occ-data.org || © 2017 Open
- Powered by [Hugo](#) and the [Kube theme](#)

OCC GOES-16 Resources

<http://edc.occ-data.org/goes16/getdata/>

Get Data

- 01 [Provisional Data](#)
- 02 [Best Effort](#)
- 03 [File Formats](#)
- 04 [Products Available & Name](#)
- [Conventions](#)
- 05 [Data Access](#)

Provisional Data

NOAA's GOES-16 satellite has not been declared operational and its data are preliminary and undergoing testing.

Best Effort

The data are being made available on a best effort basis by all parties involved. There are no guarantees the data will be available when you really need it.

NOAA would appreciate your feedback

- Are the types of data access and services provided by the BDP and Collaborators meeting your needs?
- Does the BDP approach make things easier on the user?
- Encourage communications with the Collaborators
 - Help shape the services that you need
- Seek feedback from NOAA on the BDP, NOAA data in general, and the GOES-16 data in particular
 - BDP: Ed Kearns ed.kearns@noaa.gov
 - GOES-16: Renata Lana renata.lana@noaa.gov

Summary

- NOAA is collaborating with industry through the Big Data Project CRADAs to learn how to make NOAA's full and open data **more easily and widely usable**, in a cost-effective manner.
 - GOES-16 Data are available now at BDP Collaborators' sites
 - NOAA seeks and welcomes your feedback!
- The BDP experiment is showing that modern platforms may provide:
 - **Higher Levels of Service** to the customer
 - **Reduced loads on NOAA** access systems that may reduce cost
 - **Efficient methods for data discovery and integration**
- Can applications can be developed **faster and more efficiently?**
 - Authoritative data are co-located with the processing capacity
 - Lower barriers to use for the public and small businesses?



Thank You

ed.kearns@noaa.gov

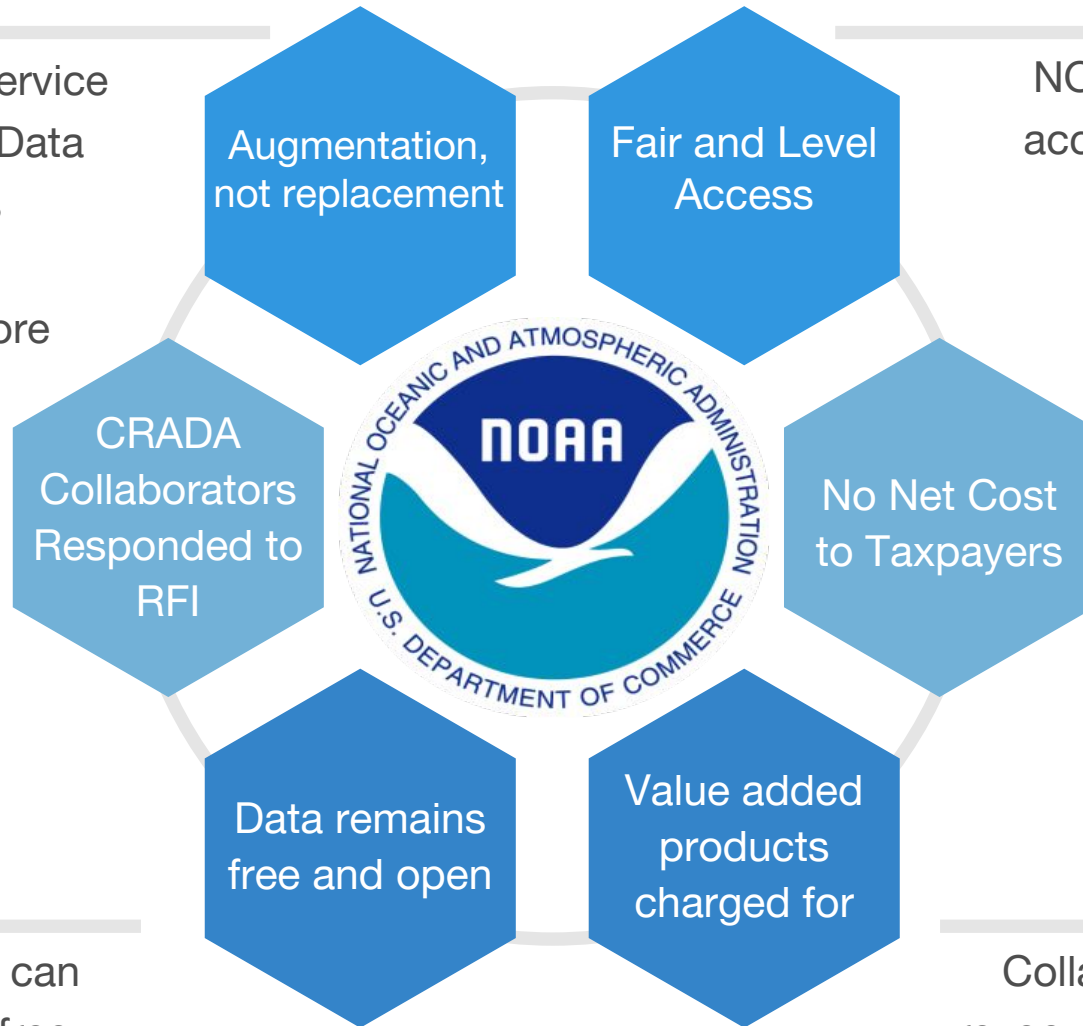
#NOAABigData

<http://www.noaa.gov/big-data-project>

All existing NOAA service outlets remain. Big Data Project (BDP) offers alternatives and advantages to explore

Collaborative Research And Development Agreement (CRADA)

Original NOAA data can be downloaded for free through collaborators. Collaborators may recover costs associated with data acquisition

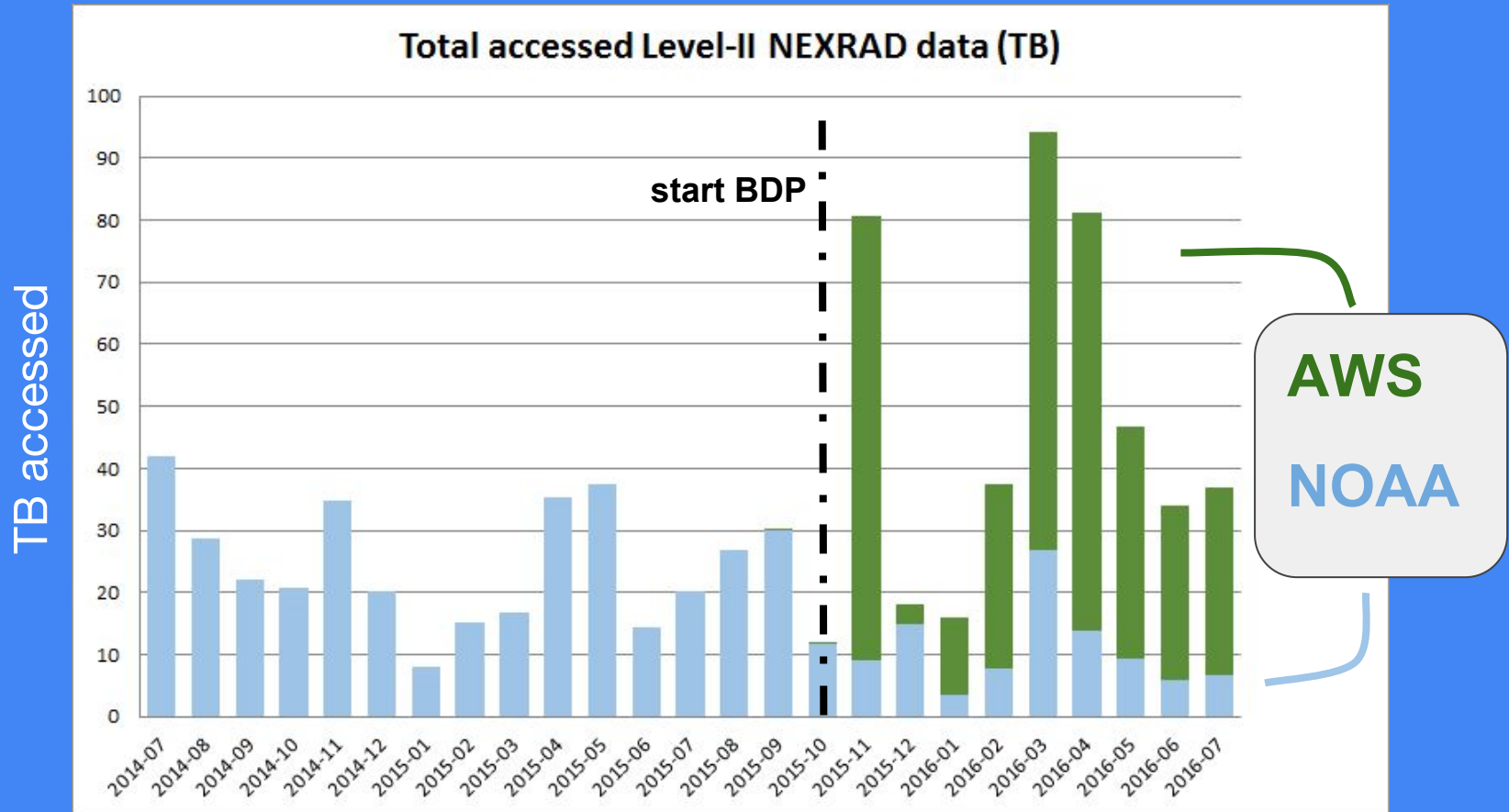


NOAA will offer equal access to the data for all collaborators

As part of the CRADA, NOAA may recover costs for new or supplemental efforts

Collaborators generate revenue when 3rd parties process the data. Collaborators may charge for value-added services and products

NEXRAD Weather Radar Data



AWS: Oct '15 <https://s3.amazonaws.com/noaa-nexrad-level2> (1991+)

OCC: Jun '16 <http://occ-data.org/NOAANEXRAD/> (2015+)

Google: June '17 <https://cloud.google.com/storage/docs/public-datasets/nexrad> (1991+)

(S. Ansari et al, 2017)